

Włodarczyk, M., Kopaczyk, J. and Kozak, M. (2020) Multilingualism in Greater Poland court records (1386-1448): tagging discourse boundaries and code-switching. *Corpora*, 15(3), pp. 273-290.
(doi: 10.3366/cor.2020.0200)

This is the author's final accepted version.

There may be differences between this version and the published version. You are advised to consult the publisher's version if you wish to cite from it.

<http://eprints.gla.ac.uk/218506/>

Deposited on: 18 June 2020

Multilingualism in Greater Poland court records (1386-1448): Tagging discourse boundaries and code-switching

Matylda Włodarczyk, Adam Mickiewicz University in Poznań
Joanna Kopaczyk, University of Glasgow
Michał Kozak, Poznań Supercomputing and Networking Center

Abstract

The report presents the digitisation project of a diplomatic edition of mediaeval land court oaths recorded in Latin and Old Polish (*Electronic Repository of Greater Poland Oaths, eROThA*, 1386-1446). We present the background, aims, design and methodology behind a small, lightly tagged specialised bilingual corpus. We also discuss the problems and limitations entrenched in the task of digitising a printed diplomatic edition into a machine-readable diplomatic edition equipped with a new interpretative layer sensitive to the switches between Latin and Old Polish. In addition to automatic annotation of code-switched items on the basis of typographic characteristics of the printed edition, flexible coding of recurrent language and discourse boundary phenomena has been introduced manually to account for linguistically ambiguous or neutral forms. The project offers a fully multilingual corpus, as well as customised Polish-only and Latin-only datasets, and enables filtered metadata searches in the online front-end. Overall, the report presents a methodology for constructing multilingual corpora in the context of legal cultures in medieval Central Europe that may be extrapolated to datasets originating in different periods and regions.

Keywords: code-switching, tagging, Old Polish, mediaeval land courts, Greater Poland

1. Introduction: Building a multilingual historical corpus

Recent corpus compilation efforts have been marked by two tendencies: on the one hand, striking a balance between representativeness and size in the spirit of big data, and, on the other, acknowledging the significance and utility of small specialised datasets. The project summarised in this paper represents the latter trend, which has been particularly distinct in historical corpus linguistics (Claridge, 2008; Kytö, 2012: 1512-1514). Historical datasets are fragmentary by nature and pose very specific research questions which may be addressed by means of specialised databases. The recent interest in historical multilingualism in particular has been fuelled by small and medium-sized specialised corpora and it is on this basis that new theoretical and methodological frameworks are being developed (see among others Schendl and Wright, 2011 eds., Nurmi et al. 2017 eds., Peikola et al. 2017 eds., Pahta et al.

¹ The paper and the free electronic database are a result of a research project funded by the National Science Centre in Poland (OPUS No. 2014/13/B/HS2/00644 [<https://projekty.ncn.gov.pl/index.php?s=11260>]).

eds. 2018). This work has mostly relied on texts written in administrative, legal and religious contexts, especially in pre-Renaissance and the Early Modern period. The characteristic feature of such texts across Europe, alongside the use of Latin as the traditional and prestigious written medium, is that they contain glimpses of the vernaculars, in the spoken domain (e.g. Ingham and Marcus, 2016 on Anglo French and Latin; Stam, 2017 on Irish and Latin), and, inevitably, code-switching (henceforth CS) from and to Latin (or more languages in some cases; e.g. Kopaczyk, 2013; Lazar, 2016).

In order to achieve systematic ways of investigating and interpreting historical multilingualism, the field is in the process of developing guidelines and digitalisation policies for the presentation of multilingual primary material. In this respect, scholars have emphasised a growing need for tagging solutions that are entrenched in broader frameworks of multilingual communication via the written medium. Such solutions have already been adopted for the languages with most extensive digital documentation, such as English (Tyrrkkö et al. 2017). In the case of the languages whose histories have not been so well documented, resources are frequently lacking to support the application of more sophisticated semi- or fully automatised tagging of language boundaries and switches. Importantly for our project, there are no dictionary resources in a fully modern digital format that could be used as reliable control corpora for automatised language recognition in mediaeval texts containing Latin and Old Polish. Although useful electronic reference works exist, none of these had been designed specifically for the purpose of studying CS between Latin and the vernacular. Compared to the existing diachronic corpora that include Old Polish (and beyond), the novelty of the *eROThA* tagging scheme lies in (1) the marking of all multilingual elements following the TEI; (2) separate marking of the linguistically ambiguous discourse boundary elements (e.g. visual diamorphs) by means of the attribute ‘source’; (3) incorporation of visual cues in the tagging of CS.

It is due to the Latin–Polish bilingualism in the legal administrative domain in the late fourteenth and early fifteenth century that the earliest extant passages in Old Polish survive within the Latin text. This bilingual practice is richly attested in the records of land courts all over the Polish-speaking territory and the source texts have been published in printed editions (e.g. Kuraszkiewicz and Wolff, 1950), but not in an electronic format. Our project concentrates on the Greater Poland land court records, since the monumental five-volume diplomatic edition of these texts by Kowalewicz and Kuraszkiewicz (1959-1981; Section 2 and 2.1 below) uses some typographic marking of CS that may be used as the basis for automatic annotation. Thus, it offers an ideal starting point to develop a methodology for constructing multilingual corpora in the context of legal cultures in medieval Central Europe.

In this paper, we present the background, aims, design and methodology behind the *Electronic Repository of Greater Poland Oaths (eROThA)* database – an electronic corpus based on the diplomatic printed edition of the Greater Poland court oaths (<http://rotha.ehum.psnc.pl>). We also discuss the problems and limitations entrenched in the task of digitising a printed diplomatic edition into a proper digital edition in the modern sense (e.g. Honkapohja et al. 2009; Vanhoutte and van den Branden, 2009). Then we present the steps taken to transform the printed edition into a digital edition (2.2). This is followed by a detailed overview of the objectives and outcomes of the project (2.3). In Section 2.4, we focus

² E.g. the following projects: *Medieval Nordic Text Archive* (MENOTA; Oslo at http://www.menota.org/EN_forside.xhtml); *Bilingualism in Medieval Ireland – Language choice as a part of intellectual culture* (Galway and Utrecht; e.g. Bisagni and Wartjens, 2007; Stam, 2017).

³ In particular, the *eFontes* dictionary (*Elektroniczny korpus łaciny średniowiecznej na ziemiach polskich*, “Electronic Corpus of Medieval Latin in Polish Lands”) and *Słownik pojęciowy języka staropolskiego* (“Old Polish Conceptual Dictionary”) proved very useful as reference works for the study of the land book data. Cf. Pastuch et al. 2018 for an overview of electronic resources for historical and contemporary Polish.

on the mark-up and annotation schemes that were used to represent CS along with the manual tagging of the so-called *triggers*, which simultaneously mark discourse and language boundaries. A short summary closes the report (Section 3).

2. Latin and the vernacular in the land court: Aims and design of the *eROThA* project

In the land courts of Greater Poland and beyond,⁴ Latin held the position of the language of the record well beyond the medieval period (Bedos-Rezak, 1996).⁵ In the oral litigation procedure, however, at least in the late fourteenth and the first three decades of the fifteenth century, the vernacular also had a crucial part to play, as showcased in the land books.⁶ As a result, in the administrative and court record, ‘Chancery bilingualism’ (Adamska, 2013) came about. In the land courts where defendants were subject to the so-called compurgation ritual, the vernacular was the vehicle for the defendant’s plea. The ritual involved giving an oath of denial (in the vernacular) in response to the accusation, or the account of the complaint witnesses (cf. Ziv and Smith, 1996: 50-51). The multilingual context in which the oath (in Latin: *rota*) is embedded constitutes the focus of the *eROThA* project. In the record, individual cases which typically begin with the administrative details provided in Latin (usually the names and provenance of the involved parties) are followed by a *rota* in Old Polish most likely produced orally in court by the defendant’s witnesses. In addition, oaths may be followed by further Latin sections concerning the dates of subsequent trials or further procedural information (Trawińska, 2014: 97).

The edited record containing over 6,330 civil and criminal cases for six locations in Greater Poland (Kopaczyk et al. 2016: 20-21; Table 1) spans six decades (1386-1446; Kowalewicz and Kuraszkiewicz, 1951-1981); although the regional and temporal representation is uneven (for details see Kopaczyk et al. 2015). The printed oaths have been employed for the study of grapho-phonemic correspondences, Old Polish syntax and morphology (e.g. Krążyńska, 2010; Słoboda, 2012),⁷ as well as legal history (Rymaszewski, 2008). According to the palaeographic analysis conducted by the editors, over 200 scribal hands were involved in record keeping. The *eROThA* project has transformed the edition into a modern open access multipurpose digital resource for linguists and historians, incorporating high resolution facsimiles. Our aim is to enhance and support the existing scholarship and inspire new lines of inquiry, also into palaeographic detail. In addition to various research topics that have already been explored based on data samples, the main interest of our project falls on the modes and constraints on CS operating on different linguistic and discourse levels, for different scribes, courts, and over time. As systematic recognition of multilingual features in these linguistically mixed texts has not been undertaken before, this projects sets

⁴ The primary data surviving for the land and town courts (*acta terrestria et castrensia*) in other Polish-speaking territories show the presence of the vernacular next to Latin well into the eighteenth century (Wąsowicz, 1975; Kulecki, 2008).

⁵ The time span of the Latin land books (180 volumes) of Greater Poland is 1396-1791 (Trawińska 2009: 345).

⁶ The editors indicate that only the earliest land books include Polish oaths (Kowalewicz and Kuraszkiewicz, 1959: 10), while in the later ones (books XV and XV for the 1440s), oaths were overwhelmingly recorded in Latin.

⁷ In Central Europe, both land law and canon law regulated ecclesiastical properties, while ‘several varieties of ‘German’ law of the inhabitants of towns and villages [...] had been created according to German models’ (Adamska, 2013: 354). The situation was similar in Hungary and the Bohemian Crowns. There may have been analogous mediaeval manor courts in the English speaking territories, albeit earlier on (the thirteenth century; Zvi and Smith, 1996: 50-51).

⁸ A selection from c. 55 land books was diplomatically edited by a palaeographer, Kowalewicz and a historical linguist Kuraszkiewicz.

⁹ More recent work has reassessed the usefulness of this material for dialect studies (see Trawińska, 2009 and references therein). The revision is in line with a growing volume of work on historical specialised registers which stresses the independence of written text and the need to study it as linguistic data in its own right, rather than as a reflection of spoken language (Jucker and Pahta eds. 2011).

out to make provisions for such elements, i.e. to enable access to language contact features via searches in an open access web-based engine. To achieve these aims, annotation of multilingual features needs to be implemented in a computer-readable and searchable version of the printed edition. However, as we have shown in our studies (Kopaczyk et al.; 2016; Włodarczyk et al. forthcoming a; Włodarczyk and Adameczyk, forthcoming b), the complexity of CS between Latin and Polish is so rich that no annotation scheme will be able to capture all its features. Instead, the compilers aim to include the tagging of CS phenomena on the highest level of linguistic and discourse organisation, as well as on the level of syntax and lexicon, albeit excluding personal and place names where CS is frequently present on the level of morphology (Kučała, 1974).

There is a growing conviction in the field of corpus building that digitisation projects based on printed diplomatic or critical editions of primary data should try to reconcile the requirements of linguistic corpora and those of digital editions (Vanhoutte and van den Branden, 2009; Honkapohja et al. 2009; Martilla, 2013). A digitised static edition should not be the ultimate outcome of such endeavours; instead researchers should aim to provide a multipurpose digital resource that involves modularity and dynamic architecture. In this way, open-endedness and combinability are achieved: these are the prerequisites for true interdisciplinarity. Honkapohja et al. state that through such an approach, which involves modular and adaptable networks of texts open to new layers of annotation, corpus compilers may be liberated ‘from the chains of “what has been edited”’ and enabled ‘to add texts from original sources with reasonable effort, effectively becoming digital editors themselves’ (2009: 467). Although the *eROThA* project does not undertake any major revision of the interpretations proposed by the printed edition, it follows the concept of an open-ended digital edition in that (1) it offers a lightly annotated corpus of data that can be used by linguists and non-linguists; (2) it provides ‘an analytical descriptive model’ (Martilla, 2013: 3) of CS on the discourse level; as well as (3) ‘an archive of research material in the form of the intermediary facsimiles and transcriptions, which may (...) be used to produce different kinds of editions in the future’ (Martilla, 2013: 4).

2.1. The unit of analysis: Oaths as individual cases

In the *eROThA* project, corpus design and data presentation formats are based on the analytical unit of discourse and the schema of representation of the communicative event, inherited from the printed edition. In the printed edition, an individual compurgation event, including an oath (a single one as a rule; in a few cases – a number of consecutive oaths formulated in connection to a single case) was selected for presentation. This decision was most likely determined by the monolingual focus: in selecting the Polish discourse element as the unit of manuscript representation, the editors aimed to foreground the earliest record of extended utterances in the vernacular.¹⁰ Thus, oaths were spotted in the books and paired with Latin introductions, thus rendering two discourse elements (the Latin and the Polish) that constitute a single ‘unit’, or one oath-taking (compurgation) event. The Polish components of such an event were diplomatically transcribed and then enhanced with editorial transcription – a quasi-translation into standardised Old Polish (achieved mostly by reducing spelling and morphological variation) (see Kopaczyk et al. 2016: 22; Figure 1 for an illustration of the presentation mode in the edition). In terms of organising the records, Kowalewicz and Kuraszkiewicz used the division into scribal hands in each individual location. In effect, the

¹⁰ This is the most likely explanation, which goes hand in hand with monolingual ideologies and the myth of national languages prevalent in linguistic and literary studies of the twentieth century, cf. Tyler (ed. 2011).

chronology and the ordering of individual books, or even of the leaves within the books, are not consistently followed, as precedence is given to the grouping by scribal hand.

Importantly for our purposes, a unit of presentation is equipped with information about the source, date of origin and scribe. These features have been transformed into the metadata used to describe the oaths in the *eROThA* database. Extracting the metadata automatically was not a straightforward process, as relevant information was not recorded in a consistent manner in the printed edition. For example, the information on scribal hands was only given once in the main text: in a separate paragraph (heading) followed by further units by this hand. Thus, the *eROThA* metadata entry had to be constructed out of the relevant pieces of information in a carefully designed automatic selection process. As a result, each entry consists of the individual unit number (reference to the printed edition), the location of the court, date of the record, source information (book and leaf number) and scribal hand.

2.2. From the printed edition to the TEI P5 scheme

In the first step, the edited volumes were converted to text files using a standard Optical Character Recognition (OCR) process resulting in html files with adjacent css files.¹¹ This process was sensitive to the structure of the text (divisions into the Latin introduction, the Polish oath, standardised editorial text and the footnotes) and its typographic format (font size, italics, etc.). Then we implemented a chain of grammars using ANTLR 4 (Parr, 2013) and the Java programming language, transforming the OCR output into standardized TEI P5 format, encoded in UTF-8.¹² The implemented parser was designed to combine and divide the text according to the metadata and to convert an individual unit into a separate xml file in the TEI P5 scheme. A sample TEI header is presented below.

```
<tei:teiHeader>
  <tei:fileDesc>
    <tei:titleStmt>
      <tei:title>Rota 330, Księga Ziemska 4, Gniezno</tei:title>
    </tei:titleStmt>
    <tei:publicationStmt>
      <tei:publisher>
        <tei:orgName>ROThA</tei:orgName>
      </tei:publisher>
      <tei:date when="2019-02-20"/>
    </tei:publicationStmt>
    <tei:sourceDesc>
      <tei:biblStruct>
        <tei:monogr>
          <tei:title xml:lang="pol">Rota 330, Księga Ziemska 4, Karta 117, Gniezno</tei:title>
          <tei:title xml:lang="eng">Rotha 330, Land Court Book 4, Page 117, Gniezno</tei:title>
          <tei:author role="author">
            <tei:persName sameAs="Per.19">PISARZ 19</tei:persName>
          </tei:author>
          <tei:textLang mainLang="lat-med"/>
          <tei:textLang mainLang="pol-old" otherLangs="pol"/>
          <tei:imprint>
            <tei:pubPlace>Gniezno</tei:pubPlace>
            <tei:date when="1430"/>
            <tei:biblScope unit="issue">330</tei:biblScope>
            <tei:biblScope unit="volume">GnZ4</tei:biblScope>
          </tei:imprint>
        </tei:monogr>
      </tei:biblStruct>
    </tei:sourceDesc>
  </tei:fileDesc>
</tei:teiHeader>
```

¹¹ OCR and Fine Reader procedures.

¹² <http://www.tei-c.org>

```

        <tei:biblScope unit="page" from="117" to="117"/>
    </tei:imprint>
</tei:monogr>
</tei:biblStruct>
<tei:msDesc>
    <tei:msIdentifier>
        <tei:repository>ROThA</tei:repository>
        <tei:idno>Gn.330</tei:idno>
    </tei:msIdentifier>
</tei:msDesc>
</tei:sourceDesc>
</tei:fileDesc>
</tei:teiHeader>

```

Figure 1. Sample TEI header

Apart from the metadata, the printed edition also defined the regions of representation for the digital version: the Latin introduction, the Polish oath (transliteration into Old Polish) and its standardised version. The above mentioned parser has converted the three regions into three separate annotated text regions within one TEI P5 file (see 2.3 below). Rich TEI-schema tags were implemented in the Latin and Old Polish sections. In addition to marking out CS, many other features of the text, such as glosses, additions, deletions and omissions, are captured in the schema. The three text regions of Gn.330 are presented in Figure 2 below. Additionally, the parser has linked each file with a high quality photograph of the relevant manuscript leaf (or leaves), featuring a red square around the corresponding source text. The images are available in the *eROThA* portal⁶ together with the automatically generated text files which are compatible with any computational tool used for linguistic analysis (e.g. *Antconc*).

```

<tei:text>
    <tei:body>
        <tei:pb n="1" facs="Gn.330/330_1430_117_GnZ4_19.tif"/>
        <tei:div type="original">
            <tei:p xml:lang="lat-med">Induccio. Testes nobilis Katherine conthoralis nobilis Lewin de Goslini
            contra nobilem Janussium Rinarewsky: Primus dominus Mathias de Labyszyno palatinus Brestensis, sicut arbiter
            et divisor, secundus Grzymko de Jarunowo, ut arbiter et divisor, tercius Thomko de Wylczyna, qui presens fuit,
            quia consensit, quartus Johannes Goleyowsky, qui presens fuit, quintus Crczon de Yeszewo, sextus Czeszek
            venator. <tei:gap/>
            <tei:foreign xml:lang="lat-med" source="independent-trigger">Rotha huius:</tei:foreign>
        </tei:p>
        </tei:div>
        <tei:div type="transliteration">
            <tei:p xml:lang="pol-old">
                <tei:foreign xml:lang="lat-med">Arbitri et divisores debent iurare sub hac forma:</tei:foreign>
            </tei:p>
            <tei:p xml:lang="pol-old">Tako <tei:lb/> gym pomofzy etc. yakofmÿ <tei:lb/>
                <tei:del>tho czøfcz</tei:del> thego dzelczø bili <tei:lb/> gdzefz panye katharzyne czøfcz zpana
                labyfkego <tei:add>ftroni</tei:add> dzalem <tei:lb/> doftala ktorøfmÿ ye podali <tei:lb/> athø czøfcz panÿ
                katharzyna <tei:lb/> rofgranyczacz ma a nÿe pan <tei:lb/> Jacufz rinarzewfkÿ </tei:p>
            <tei:p xml:lang="pol-old">
                <tei:foreign xml:lang="lat-med">Et alii tres presentes:</tei:foreign>
            </tei:p>
            <tei:p xml:lang="pol-old">yfze przywolil <tei:lb/> gdzie fzye comu doftanye ten J<tei:ref
            target="#f1">1</tei:ref> foby granycz<tei:unclear>i</tei:unclear>. </tei:p>
            <tei:p xml:lang="pol-old">

```

⁶ The images were provided by the National Archives in Poznań, one of the project partners. See, for instance, Rota 330 from Gniezno (ID = Gn.330) is available on <https://rotha.ehum.psnc.pl/breeze/Gn.330>

```

    <tei:foreign xml:lang="lat-med">Et unus in testimonium.</tei:foreign>
  </tei:p>
  <tei:note xml:id="f1">Następnie plama atramentu, która zalała kilka liter."</tei:note>
</tei:div>
<tei:div type="transcription">
  <tei:p xml:lang="pol">Tako jim pomóż etc., jakosmy <tei:del>tę część</tei:del> tego dzielcą byli,
gdzież sie panie Katarzynie część z pana Łabiskiego strony działem dostała, którą-smy je podali; a tę część pani
Katarzyna rozgraniczać ma, a nie pan Jakusz Rynarzewski.</tei:p>
  <tei:p xml:lang="pol">
    <tei:gap/>
  </tei:p>
  <tei:p xml:lang="pol">iże przywołil, gdzie sie komu dostanie ten J... sobie graniczy</tei:p>
  <tei:p xml:lang="pol">
    <tei:gap/>
  </tei:p>
</tei:div>
</tei:body>
</tei:text>

```

Figure 2: Rich TEI-schema tags

Although the printed edition provided the basic text input and structure, the digital edition has the added value of (1) annotating multilingual passages and enabling searches either in a selected monolingual section, or in the full multilingual database; and (2) opening the primary sources to new interpretations by linking the transcriptions with high-quality manuscript images.

2.3. CS at the discourse boundary

In terms of CS tagging, the project concentrates on the level of discourse, and within individual discourse elements on clausal, phrasal and lexical switches. This kind of analysis is supported by the layout employed in the printed edition, whereby a division is consistently maintained between the Latin introduction (the witness list) and the Polish oath. Within each of the discourse chunks, the printed edition renders lower-level CS elements (into Polish or into Latin) in italics. Such formatting is used for both inter- and intrasentential CS, as well as single lexical items, and translates easily into TEI marking <foreign> in an automatised parsing process which was implemented during OCR design and correction (see the workflow below). Nevertheless, in some cases the typographic marking of CS elements was incomplete, so the parser complemented it with Apache Tika language detection.¹⁵

There are three regions in the basic unit of presentation (the individual *rota*): the Latin introduction (and the optional Latin chunks that followed an oath), the Polish oath, and the original editorial translation into Old Polish (separate TEI sections <div>s, see the example in section 2.5.). In both Latin and Old Polish <div>s the items or phrases which are code-switched (italicised in the printed edition, or recognized by Apache Tika as a different

¹⁵ Editorial comment in a footnote: 'Ink stain follows blurring a number of graphemes.' Translation: Introduction. Witnesses noble Katherine spouse of noble Lewin de Goslini against the noble Janussium Rinarewsky: First sir Mathias de Labyshyno palatinus Brestensis, as judge and divider, second Grzymko de Jarunowo as judge and divider, third Thomko de Wylczyna, who is present, who have consented, fourth Johannes Goleyowsky, who is present, fifth Crczon de Yeszewo, sixth Czeszek huntsman. Judges and dividers should swear in this way: So help them etc. as we have *this part* been dividers of what lady katherzyna received by from sir labyfki's that we have decided upon this and this part lady katarzyna should distribute not sir Jacufz rinarzewfky And all three present: as he has consented what and where his share is he is to distribute. And one in testimony.

¹⁶ <https://tika.apache.org/1.17/detection.html>

language) were automatically tagged as <foreign>. The tag is dynamic in nature, as it encodes switches from Latin into Polish and from Polish into Latin, depending on the <div>. The items at the boundary of discourse and language switch (*triggers*)¹⁶ were also tagged as <foreign>, but with the additional attribute ‘source’ pointing to the type of the trigger. The attribute has the value ‘INDEPENDENT trigger’ if it is visually independent in the manuscript and the value ‘POLISH transliteration’ if it belongs visually to the Polish oath (the text region of Polish oath is denoted as ‘transliteration’). The boundary elements which did not stand out visually from the Latin introduction were not marked, even if their semantics indicated flagging or another transitional function.¹⁷ All boundary elements, whether tagged or not, remain part of the Latin introduction <div>. This approach has allowed for creating datasets which are Latin only (excluding the CS), or Polish only, while at the same time remaining sensitive to the special nature and the dynamic positioning of the trigger.

The three regions described above (see also 2.1), and the <foreign> marked sections were used in the indexing phase of TEI documents in the SOLR search engine working in the background. In effect, the users can search within any of the three text regions, while the exclusion of the multilingual elements or triggers is an additional option they may choose. The following metadata filters may be applied: location, scribal hand and year. The entire design (the website with TEI indexing, a user-friendly display and search options) has been implemented in Java programming language with the use of Play Framework. We provide a host of download functions: from TEI files of whole collections, through individual *rota* units, to particular text layers of a given unit. Additionally, our interface allows the user to export full context search results in tab-separated value files (.tsv),¹⁸ which can be further processed by corpus linguistic tools.

2.4. Project workflow

In brief, the *eROThA* project involved the following interconnected stages:

- 1) Correction (the first round in Fine Reader training mode) of the ‘dirty’ OCR, including:
 - implementation of the parser-readable scheme
 - manual tagging of the discourse boundary element (see 2.3 above and Włodarczyk et al. 2017)
 - manual correction of the content
 - manual correction of the html scheme

Although the OCR process was conducted by means of specialised software in a learning modus and followed a relatively painstaking training period (c. 1-2 *rota* units for 50-100 batches of 5 scans), the accuracy of the recognition was not completely satisfactory. The crude OCR procedure produced errors in ca. 10%-15% of the tokens, but as the training process was dynamic, at the end of a training session (per file consisting of ca. 10-15 pages) the error level was lower and a manual correction was conducted on the residue of the incorrect tokens. Detailed analyses of the mistakes that had to be corrected manually were not performed, but the patterns observed allow indicating three factors that were most likely to

¹⁶ Triggers may be visually independent of any of the sections of the manuscript, thus constituting a special status element which may be linguistically ambiguous (cf. the notion of homophonous diamorphs in Muysken, 2000 and visual diamorphs in Wright, 2001; for a detailed discussion of the trigger element, see Włodarczyk and Adamczyk forthcoming c). Even though xml does not handle overlaps, it is still possible to preserve the dynamic nature of the trigger by deploying this special annotation.

¹⁷ The strategies of boundary marking in the Latin introduction require a thorough investigation, ideally based on a specific visual-semantic-pragmatic typology.

¹⁸ .tsv file stands for tab-separated values file - a simple text format for storing data in a tabular structure where each field is separated from the next by tab character.

have affected the accuracy. First of all, the quality of the printed source was not consistent and the pages with the lowest quality produced the least accurate outcomes. Secondly, the accuracy of the OCR process was much poorer for the region which included many special characters (i.e. the Old Polish transliteration) and better for the regions which involved transcription and spelling modernisation (Latin introduction and standardised Polish sections). Thirdly, there was a group of graphemes which were confused despite repeated training attempts: these included ‘long s’ <f> (U+017F) vs. <f>, <u> vs. <n> and <y>, <ÿ> (U+00FF) and <ÿ> (U+1E8F) and the digraphs <ni> vs. <m>, <ci> vs. <d>, vs. <h>, <nr> vs. <rn> and others. Division marks (l), brackets and hyphens were frequently not read correctly or omitted by the software. More typical OCR mistakes involved small <l> confused with the numeral <1> and capital <R> frequently recognized as . Finally, the OCR process had to be preceded by a manual division of scan pages into regions which were then ordered manually according to the requirements of the parser. In some cases, in the transfer of the corrected file to the xml format regions were curiously reordered. This was later amended manually in an xml editor.

2) Correction (the second round, in html printouts): to increase inter-coder reliability, the coders switched datasets to verify tagging decisions, content and scheme accuracy. The coders cross-checked their decisions and discussed disagreements (c. 5%). In general, the so-called borderline cases, where the visual marking of the trigger was ambiguous, were problematic. In such cases, the coders followed a strict policy of treating the trigger as independent only if it occurred in a new line or was preceded by an intentional gap that separated it from the Latin introduction.

3) Parser verification focusing on the section, language and CS scheme accuracy

4) Final correction round: minor consistency amendments

5) xml correction round: final adjustments (line breaks, word divisions, etc.) and manual correction of the TEI scheme.

2.5. Project outcomes

The digital edition of the Kowalewicz and Kuraszkiewicz *Greater Poland Court Oaths* complements and enhances the available resources for the study of mediaeval Latin-vernacular CS through up-to-date and robust methodological decisions and design. The project offers:

1) A searchable version of the oaths which is sensitive to their multilingual character. Old Polish and Latin are presented in TEI-coded sections, with internal automatic tagging of <foreign> elements within each of the sections (inter- and intrasentential CS and single lexical items).

2) Metadata filtering in online searches. The user is able to define the region and language selection, exclude either of the languages, or focus on CS elements exclusively.

3) Flexible coding of the discourse boundary material.

4) High-quality images of the original manuscripts. In this way, the primary material is open to new research questions which can be asked of the material according to various criteria, such as:

- the unit of presentation/analysis
- scribal hand
- transliteration (e.g. treatment of abbreviated items; capitalisation) and font choice
- transcription
- modernised version

6) A clear presentation of the editorial divisions into regions and languages in the manuscript scans and in the TEI coding (cf. also section and Figure 2 above). A thin line separates each region in the image and different font colours are used to display different languages. For other TEI elements, clear formatting was used, like underline, strike-throughs, superscripting, etc.

7) Automatic export of editorial regions and search results to text files compatible with basic linguistic software such as *AntConc* or *Wordsmith*.¹⁹

The database was checked for compatibility with *Antconc* in the txt and xml format and no coding compatibility issues occurred. For the former xml format, searches in the tagged and untagged display options were conducted. The tagging proved overall compatible with *Antconc*, although search strings had to be supported by the use of wildcards. In forthcoming papers, which present a more detailed overviews of code-switched elements, we have employed the searches for the <foreign> tag by means of the string: *<tei:foreign xml:lang=*. As a result, all the elements tagged automatically (i.e. the italicized items from the edition) and manually (triggers) could have been extracted for the purpose of further categorization. The simple sorting option (in 5, 6, 7 to the left) has allowed ordering the <foreign> strings according to the section in which these occurred and the direction of the switches (<tei:foreign xml:lang="lat-med"> for Latin in the Old Polish text and <tei:foreign xml:lang="pol-old"> for Old Polish in the Latin text). Unsorted search results are presented in Figure 3. The option file view (activated from Concordance or of file list) in Antconc and renders TEI schemes identical to those presented in Figures 1 and 2.

¹⁹ The application of WordSmith tools (word list and concordance) resulted in some compatibility issues affecting character encoding. These issues were partially resolved by means of adjustments in language setting features. However, as the *AntConc* software involved no such issues, it is recommended for corpus-analytic work on the *eROThA* files.

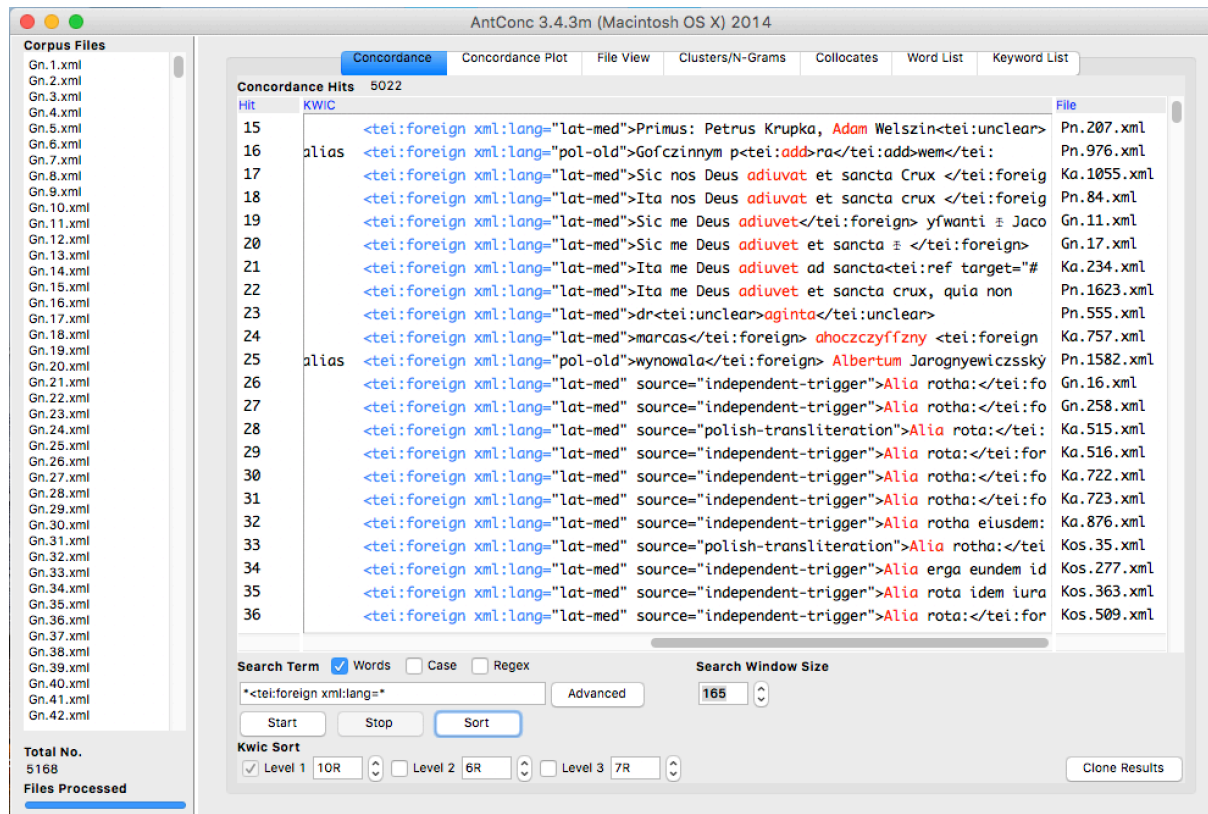


Figure 3: Antconc search sample

3. Summary

The significance of the coexistence of different languages in written texts of the past has recently come to the fore not only in historical linguistics, but also in communication studies (Garrison et al. eds. 2013). In particular, the interfaces of Latin and the vernaculars in mediaeval and Early Modern Europe have become an attractive new research area. Projects involving both the so far unexplored and already known data have been undertaken; many of them exploit administrative and legal genres. It is in these contexts that the position of Latin remained strong, or even dominant, well into the Early Modern era, but the vernaculars also have a place in these communicative contexts. The Latin–vernacular bi- and multilingualism is attested all over Europe. Various research projects thus involve different languages and locations but the scribes, usually professional clerks trained in prominent European centres of learning, remain a unifying feature. However, the degree to which the coexistence of Latin and the vernaculars in different texts shows similarities or differences across space, time and genres remains to be explored.

Overall, not unlike other corpora of mediaeval specialised texts that aim to shed light on the underlying interplay of Latin and the vernaculars, i.e. the extent of multilingualism and switches, the digitised edition of the land court oaths from Greater Poland expands our knowledge of language variation, language contact and language change. As the digitised *eROThA* provides easy access to high-quality manuscripts, one direction of study that may bring further advances in this respect is a new, technology-assisted scribal hand analysis. Such an endeavour might further add to our knowledge on the role of bi-/multilingual users in the conventionalisation of specialised registers and genres, as well as more generally, in language change, the diffusion of change in multilingual settings and the role of multilingual scribes as agents of variation and change.

Primary sources

Elektroniczny Korpus Łaciny Średniowiecznej na Ziemiach Polskich, (*Electronic Corpus of Medieval Latin in Polish Lands*), http://scriptores.ijp-pan.krakow.pl/fontes/efontes/run.cgi/first_form

Kowalewicz, H. and W. Kuraszkiewicz (eds.). 1959-1981. *Wielkopolskie Roty Sądowe XIV–XV Wieku* [The Greater Poland Court Aaths of the 14th-15th Century], vol. 1, Roty poznańskie [The Poznań oaths], vol. 2, Roty pyzdrowskie [The Pyzdry oaths], vol. 3, Roty kościańskie [The Kościan oaths], vol. 4, Roty kaliskie [The Kalisz oaths], vol. 5, A, Roty gnieźnieńskie [The Gniezno oaths], B, Roty konińskie [The Konin oaths]. Warszawa, Poznań, Wrocław, Kraków and Gdańsk: Państwowe Wydawnictwo Naukowe.

Słownik Pojęciowy Języka Staropolskiego (*Old Polish Conceptual Dictionary*)
<http://spjs.ijp.pan.pl/spjs/strona/opisProjektu>

References

Adamska, A. 2013. 'Latin and three vernaculars in East Central Europe from the point of view of the history of social communication' in M. Garrison, A. Órban and M. Mostert (eds.) *Spoken and Written Language. Relations between Latin and the Vernacular Languages in the Earlier Middle Ages*, pp. 325-364. Turnhout: Brepols.

Bedos-Rezak, B. 1996. 'Secular administration' in F.A.C. Mantello and A.G. Rigg (eds.) *Medieval Latin. An Introduction and Bibliographical Guide*, pp. 195-229. Washington: The Catholic University of America Press.

Bisagni, J., and I. Warntjes. 2007. 'Latin and Old Irish in the Munich Computus: A reassessment and further evidence', *Ériu* 57, pp. 1–33.

Claridge, C. 2008. 'Historical corpora', in A. Lüdeling and M. Kytö (eds.) *Corpus Linguistics: An International Handbook. Vol. 1.*, pp. 242–259. Berlin/New York: Mouton de Gruyter.

Garrison, M., A. Órban, and M. Mostert (eds.). 2013. *Spoken and Written Language: Relations between Latin and the Vernacular Languages in the Earlier Middle Ages*. Turnhout: Brepols.

Honkapohja, A., S. Kaislaniemi, and V. Marttila. 2009. 'Digital editions for corpus linguistics: Representing manuscript reality in electronic corpora' in A. H. Jucker, D. Schreier and M. Hundt (eds.), *Corpora: Pragmatics and discourse*, pp. 451-475. Amsterdam: Rodopi.

Ingham, R., and I. Marcus. 2016. 'Vernacular bilingualism in professional spaces, 1200 to 1400' in A. Classen (ed.) *Multilingualism in the Middle Ages and Early Modern Age: Communication and Miscommunication in the Premodern World*, pp. 145-165. Berlin: De Gruyter.

Kopaczyk, J. 2013. 'Code-switching in the records of a Scottish Brotherhood in early modern Poland-Lithuania', *Poznań Studies in Contemporary Linguistics* 49 (3), pp. 281-319.

Kopaczky, J., M. Włodarczyk, and E. Adamczyk. 2016. 'Medieval multilingualism in Poland: Creating a corpus of Greater Poland Court Oaths (ROThA)' *Poznań Studies in Contemporary Linguistics* 51 (3), pp. 9-35.

Krążyńska, Z. 2010. 'Średniowieczne techniki rozbudowywania zdań (na przykładzie wielkopolskich rot sądowych)' [Medieval techniques of syntactic elaboration (based on the Greater Poland court oaths)], *Kwartalnik Językoznawczy* (3-4), pp. 1-16.

Kucała, M. 1974. 'Łacińska fleksja rzeczowników polskich w tekstach średniowiecznych' [Latin inflections of Polish nouns in medieval texts] in J. Kuryłowicz and J. Safarewicz (eds.) *Studia Indoeuropejskie* [Indo-European studies], pp. 91-96. Wrocław: Ossolineum.

Kulecki, M. 2008. 'Zespoły ksiąg sądów szlacheckich I instancji – wprowadzenie' in D. Lewandowska (ed.) *Archiwum Główne Akt Dawnych w Warszawie. Informator o zasobie archiwalnym*, pp. 95-101. Warszawa: Archiwum Główne Akt Dawnych.

Kuraszkiewicz, W., and A. Wolff. 1950. *Zapiski i Roty Polskie XV-XVI Wieku z Ksiąg Sądowych Ziemi Warszawskiej*. Kraków: Prace Komisji Językowej PAU nr 36.

Kytö, M. 2012. 'New perspectives, theories and methods: Corpus linguistics' in A. Bergs and L. Brinton (eds.) *English Historical Linguistics: An International Handbook*. Vol. 2, pp. 1509-31. Berlin: De Gruyter Mouton.

Lazar, M. 2016. 'Grenzüberschreitungen: Stadtbücher aus der Westslovakei, Schlesien und Kleinpolen und Interpretationen ihrer Mehrsprachigkeit' Paper presented at *Workshop Fontes Iuris Lusatiae Superioris Vetustissimi*. Kraków 2016.

Muysken, P. 2000. *Bilingual Speech: A Typology of Code-Mixing*. Cambridge: Cambridge University Press.

Nurmi, A., T. Rütten, and P. Pahta (eds.). 2017. *Challenging the Myth of Monolingual Corpora: Multilingualism in English Corpora*. Amsterdam: Brill/ Rodopi.

Pahta, P., and A. H. Jucker (eds.). 2011. *Communicating Early English Manuscripts*. Cambridge: Cambridge University Press.

Pahta, P., J. Skaffari, and L. Wright (eds.) 2018. *Multilingual Practices in Language History. English and Beyond*. Berlin: De Gruyter.

Parr, T. 2013. *The Definitive ANTLR 4 Reference*. San Francisco: The Pragmatic Bookshelf, Raleigh.

Pastuch, M., B. Duda, K. Lisczyk, B. Mitrenga, J. Przyklenk, and K. Sujkowska-Sobisz. 2018. 'Digital Humanities in Poland from the perspective of the historical linguist of the Polish language: Achievements, needs, demands', *Digital Scholarship in the Humanities* 33 (4), pp. 857-873. <https://doi.org/10.1093/llc/fqy008>

Peikola, M., A. Mäkilähde, H. Salmi, M.-L. Varila, and J. Skaffari (eds.). 2017. *Verbal and Visual Communication in Early English Texts (Utrecht Studies in Medieval Literacy 37)*. Turnhout: Brepols.

Rymaszewski, Z. 2008. *Z Badań nad Organizacją Sądów Prawa Polskiego w Średniowieczu. Woźny Sądowy*. [From the Research on the Organisation of Polish Law Courts in the Middle Ages. Court Usher]. Warszawa: Akademia Leona Koźmińskiego.

Schendl, H., and L. Wright (eds.). 2011. *Code-switching in Early English*. Berlin: Walter de Gruyter.

Słoboda, A. 2012. *Liczebnik w Grupie Nominalnej Średniowiecznej Polszczyzny. Semantyka i Składnia* [Numeral in the Nominal Group of Medieval Polish. Semantics and Syntax.] Poznań: Wydawnictwo Rys.

Stam, N. 2017. *A Typology of Code-switching in the Commentary to the Félice Óengusso*. Utrecht: LOT publications.

Trawińska, M. 2009. 'Cechy dialektalne wielkopolskich rot sądowych w świetle badań nad rękopisem poznańskiej księgi ziemskiej' [Dialect features of the Greater Poland court oaths: The analysis of the Poznań municipal book manuscript], *Prace Filologiczne* LVI, pp. 345-360.

Trawińska, M. 2014. *Rękopis Najstarszej Poznańskiej Księgi Ziemskiej (1396-1400)*. [Manuscript of the Oldest Poznań Land Book (1386-1400)]. Warszawa and Poznań: Wydawnictwo Rys.

Tyler, E. M. (ed.) 2011. *Conceptualizing Multilingualism in England c.800-1250*. Utrecht: Brepols.

Tyrkkö, J., A. Nurmi, and I. Tuominen, J. 2017. 'Semi-automatic discovery of code-switching from English historical corpora: Methods and challenges' in A. Nurmi, T. Rütten and P. Pahta (eds.) *Challenging the Myth of Monolingual Corpora: Multilingualism in English Corpora*, pp. 172-199. Leiden: Brill.

Wąsowicz M. 1975. 'Księgi ziemskie i grodzkie (acta terrestria et castrensia) — wprowadzenie ogólne' in J. Karwasińska (ed.) *Archiwum Główne Akt Dawnych w Warszawie. Przewodnik po zespołach. I. Archiwa dawnej Rzeczypospolitej*, pp. 147–155. Warszawa: Archiwum Główne Akt Dawnych.

Wright, L. 2011. 'On variation in medieval mixed-language business writing' in H. Schendl and L. Wright (eds.) *Code-Switching in Early English*, pp. 191–218. Berlin/Boston: De Gruyter Mouton.

Włodarczyk, M., E. Adamczyk, and O. Makarova. Forthcoming a. 'Code-switching and literalisation in provincial court books (*libri terrestres*): Evidence from the Electronic Repository of Greater Poland Oaths (1386-1446)', in M. Lazar, and W. Carls (eds.), *Das Sächsisch-Magdeburgische Recht als Kulturelles Bindeglied zwischen den Rechtsordnungen Ost- und Mitteleuropas. Bestandsaufnahme und Perspektiven der Forschung* (Reihe „IVS SAXONICO-MAGDEBURGENSE IN ORIENTE“ 8), Berlin: Mouton de Gruyter.

Włodarczyk, M., and E. Adamczyk. Forthcoming b. 'Constraints on embedded multilingual practices in the Electronic Repository of Greater Poland Oaths (1386-1446)'.

Włodarczyk, M., and E. Adamczyk. Forthcoming c. 'Metalinguistic and visual cues to the co-occurrence of Latin and Old Polish in the *Electronic Repository of Greater Poland Oaths*, 1386-1446 (eROThA)', in M. Włodarczyk, J. Tyrkkö, J. Kopaczyk, and E. Adamczyk (eds.), *Multilingualism Meets Multimodality: Historical and Modern Contexts*.

Zvi, R., and R. M. Smith. 1996. 'The origins of the English manorial court roles as the written record' in T. Zvi and R. M. Smith (eds.) *Medieval Society and the Manor Court*, pp. 36-68. Oxford: Clarendon Press.